Are Larger Language Models Better at Disambiguation?

People experience processing difficulties when they encounter a continuation of a sentence that conforms to the less likely interpretation of the previously ambiguous syntactic structure. This phenomenon, often realized as increase in reading time, is called the garden path effect [5] and can be explained by surprisal theory of sentence processing [7, 10]. Linguistic analysis of neural language models suggests that pre-trained language models capture the syntax of natural languages [8] and represent the incremental syntactic processing states similar to those of humans [11, 6]. Through surprisal theory, previous studies observed garden path effects in autoregressive language models where the language models predicted the increase in reading time in the disambiguation region [14, 6]. However, it is unclear whether autoregressive language models finally resolve the temporary syntactic ambiguity after being exposed to the disambiguating text and whether model capacity impacts how well the model disambiguates.

To probe for the incremental syntactic representations of a language model, we analyzed the completions generated by the language model for a set of garden-path prefixes where the model's completions are indicative of its incremental syntactic representation, i.e., whether the model disambiguates the given prefix. The corpus includes 30 sentences that were adapted from an experimental study[1]. Each sentence contains a structural ambiguity between a relative clause interpretation and a complement clause interpretation. An example is *The householder told the builder that he had arranged to pay that the bill was fair.* Before encountering *that the bill*, the part of the sentence, *that he had arranged to pay*, can be interpreted either as a relative clause that modifies *the builder* or as the complement of the verb *told*. After encountering *that the bill*, the reader will realize that the relative clause interpretation is the only consistent interpretation. Another example is *The musician told the guitarist that he was impressed by that the play was appalling*. Note that it has a similar structure but is in a different context.

If the language model successfully removes the inconsistent interpretation from its representation, it would recognize *that he had arranged to pay* as a relative clause and recognize *the bill* as the start of a complement clause. If this is the case, *the bill* would be the argument of some predicate in the model's completion of the prefix (case 1). However, if *the bill* is not an argument of any predicate in the content clause (case 2), as in the ungrammatical completion, *The householder told the builder that he had arranged to pay that the bill in two weeks*, the model did not successfully disambiguate the prefix. These two scenarios can be discriminated by a dependency parser, e.g., the spaCy dependency parser [9] by analyzing the structure of the content clause (i.e., the part of the completed sentence after the second *that*). We can then evaluate how well a language model disambiguates this type of garden path sentence by sampling multiple completions for each prefix and calculating the percentage of grammatical completions generated by the model for the 30 garden path prefixes. Figure 1 shows the proportion of grammatical completions for language models of different sizes [12, 16, 15, 4, 3, 2, 13]. The line fit shows a trend that larger models are worse at generating grammatical completions so they are worse at disambiguating this type of garden path sentences. This contradicts the intuition that larger models has better linguistic capabilities and casts doubts on the hypothesis that language models maintain explicit syntactic structures during their incremental processing of language.

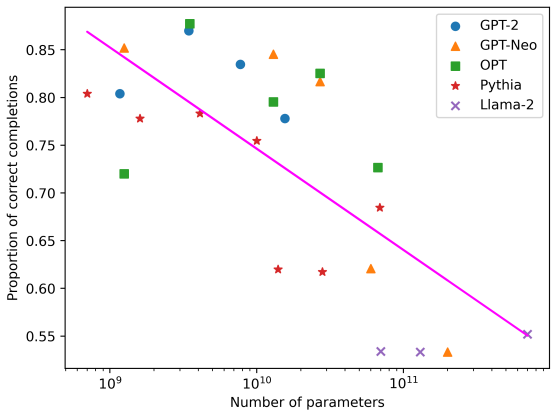

Figure 1: The proportion of grammatical completions for models of different sizes. It shows a log-linear relationship between the model's size and proportion of grammatical completions. A permutation test (10,000 samples) shows that the negative slope is significant ($p < 0.005$).

References

[1] G. T. M. Altmann, A. Garnham, and Y. Dennis. Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, 31:685–712, 1992.

[2] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[3] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

[4] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *If you use this software, please cite it using these metadata*, 58(2), 2021.

[5] L. Frazier and K. Rayner. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14:178–210, 1982.

[6] R. Futrell, E. Wilcox, T. Morita, P. Qian, M. Ballesteros, and R. Levy. Neural language models as psycholinguistic subjects: Representations of syntactic state. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] J. Hale. A probabilistic earley parser as a psycholinguistic model. In *North American Chapter of the Association for Computational Linguistics*, 2001.

[8] J. Hewitt and C. D. Manning. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[9] M. Honnibal and M. Johnson. An improved non-monotonic transition system for dependency parsing. In *Conference on Empirical Methods in Natural Language Processing*, 2015.

[10] R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177, 2008.

[11] R. Marvin and T. Linzen. Targeted syntactic evaluation of language models. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[13] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[14] M. van Schijndel and T. Linzen. Modeling garden path effects without explicit hierarchical syntax. *Cognitive Science*, 2018.

[15] B. Wang and A. Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021.

[16] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models, 2022. *URL https://arxiv. org/abs/2205.01068*, 3:19–0, 2023.